2019 CANARIE RDM Workshop

January 23, 2019

Notes and Resources

Final Version, March 7, 2019
Mark Leggott and Laura Gerlitz

DOI: 10.5281/zenodo.2584274

Introduction

This was the first Workshop for recipients of funding from CANARIE's RDM Program. A total of nine projects were funded, and all projects were represented at the Workshop. The Agenda and Background Document are also available.

Welcome and Goals

1. Looking for any synergies/collaborations.
2. All the RDM platforms/services are different but looking to see how they fit together, and if they can be integrated into a national framework.
3. CANARIE has previously added elements of collaboration into the Calls but that was not done this time as it was the initial call for the RDM Program.

Presentations

1. Welcome and Goals: https://doi.org/10.5281/zenodo.2556764
2. Research Software: Lessons Learned Over 12 Years: https://doi.org/10.5281/zenodo.2556917
3. Canadian Health Omics Repository, Distributed (CHORD): https://doi.org/10.5281/zenodo.2556920
4. Dataverse for the Canadian Research Community: Developing reusable and scalable tools for data deposit, curation, and sharing: https://doi.org/10.5281/zenodo.2557705
5. DuraCloud: https://doi.org/10.5281/zenodo.2556925
6. FAIR Repository for Annotations, Corpora and Schemas: https://doi.org/10.5281/zenodo.2556933
7. Federated Geospatial Data Discovery for Canada – Geodisy: https://zenodo.org/record/2556952
8. Making Identifiers Necessary to Track Evolving Data (MINTED) – A Brief Overview: https://doi.org/10.5281/zenodo.2556959
9. Active Research Data Management Tools: Radiam: https://doi.org/10.5281/zenodo.2556772
10. A Research Lifecycle Approach using Islandora 8 Overview: https://doi.org/10.5281/zenodo.2556963

11. Research Portal for Secure Data Discovery, Access and Collaboration: https://doi.org/10.5281/zenodo.2555384

Summary of Workshop Discussion

4. Priorities from the Discussion
    a. ID core infrastructure (ie. Datacite) should have strong governance and sustainable funding.
    b. A consistent set of services and standards supporting consent requirements.
    c. Sustainable funding is critical for DRI created by the community, but a key issue is how to determine what government should fund?
    d. A strategy for supporting and developing HQP in the ecosystem.
        i. There are issues with recruiting/retaining experts for projects.
        ii. Mention of the FutureSkills program as an opportunity for training and HQP development in data science.
        iii. College graduates may be an untapped resource.
    e. Need to provide a more inclusive context to consider indigenous issues.
5. Feedback and Ideas for Future Workshops
    a. Useful to hear more about fellow grant recipients and where they're going.
    b. Nudges things towards more harmonized approaches, opportunity to make connections with people moving in same direction.
    c. Useful to have people here who aren't PIs etc. but part of the project teams.
    d. Useful to have the Workshop and NDSF Summit back to back.
    e. Appreciated hearing about other projects, able to see connections between the projects very clearly.
    f. The longer a window to accomplish a project, the greater the possibility of collaboration.
        i. RDM submission process has been very tight.
        ii. Workshop has been useful to get conversations going, not sure how much the projects can be steered towards new collaborations coming out of workshop.
        iii. Timing for workshop - better at the beginning of the projects; this timing has been good.
    g. Sharing of knowledge of process and strategies could be fruitful.
    h. For future RDM software projects: CANARIE could consider an EOI round, give groups access to EOIs; this may work better for this group than the software groups.
    i. Once you do the design of the project and sign a contract you're kind of locked into it.
    j. Better idea of how the NDS might be better interconnected.
    k. NDSF context: encourage competitive approach but in a broader community context.
    l. Surfacing best practices out of these practices - services people could leverage.

6. Thematic options for future workshops.
    a. Privacy, access control.
    b. RDM lifecycle.
    c. Data storage.
    d. Relationship and dynamics between data sources and data aggregators.
    e. Developing/advancing standards of metadata.
    f. Structure future workshop appropriately: possible: 2-3 tracks in a given context, but would lose cross-pollination.
    g. Hackfest would be a good idea - consider lottery to pick topics to focus on.
    h. Policy context may be important.
    i. International scene - awareness of landscape based on a topic.
    j. Small group of people to facilitate the continuation of dialogue.

Notes From Presentations

Research Software Program

7. Research Software: Lessons Learned over 12 Years
8. Why are we doing research software in the first place?
    a. First 5 yrs: funded researchers to hire team to build software to do research (2007-2012)
        i. Little interaction between teams
        ii. Retrospective at end: commonality between research platforms, even across disciplines!
    b. 2013-2015: service model introduced.
        i. Asked funded community to find useful/common functionality within platforms, extract it, and offer as a software service.
        ii. Uptake for these services wasn't as good as thought, as projects that were funded were focused more on the research than the software, so the software met their specific needs.
    c. 2016: effort to promote to greater software reuse
        i. Asked for expressions of interest, shared them with all the groups to see if they could make more generic software.
        ii. We learned that people felt their platforms were too specific to research projects to be used by someone else.
    d. 2017: evolution of Call to encourage reuse.
        i. Last 2 Calls: funded pre-existing platforms to adapt and bring on new users, plus maintenance funding.
        ii. This Call vs from-scratch platform development: more efficient use of money over longer period of time.
        iii. Generally small # of users per platform; wanted to look for ways to fund more researchers.
    e. 2018-2020: HUBZero pilot.

        i. CANARIE partnered with pilot research teams to adapt HUBzero to meet their research needs (storing datasets and share data with collaborators).
    f. 2018-2020: Local Support at institutional level pilot.
        i. Funded local support teams to provide application-level support for software used for research.
        ii. Can use pre-existing software, more sustainable, follows FAIR principles.
        iii. National scalability - teams at other institutions with different expertise can come together.
    g. How does this apply to RDM?
        i. Thoughts:
            1. Easier to do early collaboration than later in the development cycle.
            2. More use cases - easier to integrate software and make generic software more of a priority.
            3. Big impact on users who don't have solutions today.
            4. May facilitate scalability, with respect to hardware, software and HQP.

RDC and RDM

9. Genesis of the RDM Program and National Data Services Framework
10. Efforts were made to ensure good synergy between broad stakeholder community and those in the RDM program, as well as intersections at events like the NDSF summit.
11. One of the main focuses of the call is FAIR
    a. There have been a few international funders working to embed FAIR into funding calls, but it is still rare.
    b. FAIR Principles a more "acceptable" than "open data".
12. CANARIE is interested in getting feedback on the integration of the FAIR Principles into the Calls:
    a. What does FAIR mean to the successful applicants in having had to integrate FAIR into different stages?
    b. Has it been easy or hard?
    c. How do we reinforce/clarify/highlight the role FAIR plays in future funding calls?
13. The second key component of the Call was the National Data Services Framework (NDSF). The NDSF provided a way to stimulate conversation around national services, whether that be standards, protocols, APIs, or actual platforms/services/infrastructure.
14. The Map of funded RDMp Projects is designed to provide detail and high level high level information in a single diagram - CANARIE would like to know if the Map provides useful information.

CHORD

15. Canadian Health OMICS Repository, Distributed (CHORD)
16. Background on genomics and epigenomics.

17. Genomics (meta)data includes donor info, sample info, experiment info, analysis info.
18. How to share genomic datasets and ensure participant protection?
    a. For example, epigenomic data can show environmental conditions and life habits: this is very personal and needs to be protected properly.
    b. Many ways of protecting privacy, e.g. controlled access repositories (EGA).
        i. Obtaining large datasets from these repositories can be tough for smaller institutions who may not have the resources needed for downloading and analysis.
19. Global Alliance for Genomics and Health (GA4GH) trying to establish standards to share data more efficiently while protecting privacy.
    a. Driver projects: trying to make and implement standards to make data available in different ways.
    b. The CanDIG project is one of these.
        i. Nodes at different locations where data is stored, API to tell what data is available at other sites (analysis bundles) without having to transfer everything from one location to another.
20. CanDIG CHORD
    a. Add/enhance modules in CanDIG infrastructure to make Canadian data more accessible to everyone in health research.
    b. 4 deliverables:
        i. Data Publishing: offer ways to upload data in CHORD infrastructure; include data ingestion mechanics which respects established standards; includes PIDs.
        ii. Findability and access: discovery portal, GENAP portal will be adapted to enable data discovery in core infrastructure; CAF integration.
        iii. Privacy-Preserving Ruse of Data: data use ontologies (clarify what can and cannot be done with dataset); differential privacy; authorization engine - who uses my data and how?
        iv. Expand CanDIG for better Health Data Support
    c. What is the driver for depositing and sharing data?
        i. Depositing your data is good visibility for your project, and data availability is increasingly a requirement for funding agencies
    d. Comment that the ONC framework also has fine-grained permission-based access, and a sandbox for researchers to run analysis on data not on their own machine
    e. What kinds of PIDs? Problems with DOIs as they would point to GENAP portal not the data.
        i. You will always be able to access the metadata, including information on where to get the data.
    f. Restricted access to data - can you speak more to providing researchers access to sensitive data?
        i. Working on mechanisms to federate permission controls to owners of data; establishing universal roles that define access.

Dataverse

21. Dataverse was developed at Harvard, and represents a very active and growing community.
22. Scholars Portal Dataverse: shared service of OCUL since 2012:
    a. Hosted in Canada
    b. Bilingual functionality
    c. Provides institutional branding
23. CANARIE grant is facilitating improvements to platform, including:
    a. features to make DV a more scalable national option
    b. Scalability
        i. Improve system scalability and provide more robust platform
        ii. Connect to existing Canadian storage environments
        iii. Ontario Library Research Cloud, built for preservation - cloud storage can be scaled more easily and taps into robust pre-existing infrastructure
        iv. Support Globus endpoints: File access could be mediated outside platform
        v. Large file upload could bypass DV
        vi. New cloud storage could handle up to 10GB
    c. Authentication
        i. ORCID: more fully integrate with platform
        ii. Advantage: no need to store local IP info in application; can be managed by schools themselves
        iii. Data curation:
            1. Tool: allows users to adopt best practices and standards in their field; would be a modular application
    d. Prototype
        i. Tabular data ingest tool.
        ii. Can create DDI XML file, maybe codebooks, that can be used in DV.
        iii. Usability testing for tool.
    e. Q: Globus endpoints: supported by CC, how long will it be funded in the long term if no longer provided by Compute Canada?
        i. Lots can be done without a subscription; core data transfer functionality is free.
        ii. Still exploring connection between Globus, DV, CC; making sure endpoints are trusted; working on relationships with these service providers.
    f. Have to think about integrating Shibboleth (ORCID too) as not every school has it.
        i. Good example where CANARIE can facilitate conversation.

DuraCloud

24. Concerned with data preservation throughout the RDM lifecycle.
    a. Data stored in a variety of places.
    b. Digital preservation a risk management exercise; good to diversify risk profile to mitigate loss of data.
    c. Economies of scale leveraged via consortium involvement in digital preservation.
25. Scholars Portal and COPPUL had both developed their own digital preservation solutions; trying to support across respective platforms.
    a. Duraspace brought into project with the Duracloud service.
26. FAIR intersections:
    a. Accessible: via standard protocols, metadata is accessible on its own (treated like data and preserved).
    b. Reusable: tracking provenance of data in its lifespan, adherence to community standards.
    c. NDSF intersections
        i. What requirements will be put forward by TC3 policy; not all repositories are well suited to long term dig pres of data; a lot of opportunity for DuraCloud to provide a way to engage with repositories through digital preservation expertise.
    d. Status
        i. In process of setting up test instance as-is; creating storage connectors for SWIFT and LOCKSS.
        ii. Breaking dependencies on AWS - mostly done, migrating parts off of Amazon.
    e. Collaboration:
        i. Can leverage past connectors with DuraCloud (e.g. Fedora repositories) as they move forward.
        ii. Any Fedora-based repository could be a pilot.
        iii. CANARIE HUBzero instance also interesting.
    f. Does architecture allow for federation?
        i. Moving in opposite direction of federation, although leverages federated preservation storage.
        ii. Also elements of policy brokerage in DuraCloud are important.
    g. Any other preservation service providers you'd want to bring in?
        i. Would want to work with any appropriate storage services, having initial conversations with SciNet.
    h. Is there a discovery system that also allows metadata to be searched?
        i. Trying to stay out of that conversation. This is intended for preservation purposes, shouldn't be end users coming and downloading this data, should be repository and service managers. Discovery more on access repository side.

    i. Challenge with sensitive data - often recommendations to destroy data after research; time is right for a more nuanced conversation about security practices for data? What about withdrawal of consent and need to remove data from archive?
        i. Data associated with withdrawal of consent shouldn't be preserved; this is about mandates to preserve data that needs it. Value thinking about expiry dates, deaccessioning.

FRACS

27. Project hopes to bridge the gap between industry and university stakeholders.
28. Many research projects have unstructured data in different media
29. Trying to train learning modules on these datasets; need supervised learning.
30. Annotated dataset serves as gold standard for machine learning algorithms
31. Problem: a lot of projects, researchers using their own repositories, not following FAIR principles.
32. Not a lot of universal tools for annotating data; many are domain-specific.
    a. CRIM and partners are long term CANARIE funded project members via Vesta and PACTE.
33. FAIR
    a. DOIs: at first thought about having only DOIs for datasets, decided on associating them with distributions of datasets.
    b. Applying metadata to several formats in every layer (repository, catalog, dataset, distribution).
    c. PIDs for metadata and datasets.
    d. Public metadata published on web and indexable:
        i. Trying to work with harvesters like FRDR and systems like DataCite.
34. Combines live storage engine and ability to generate snapshots for public.
35. Annotating research tool fits into any platform, lots of intersection; modular annotation framework hasn't really been developed.
36. RDA meeting in Botswana - birds of a feather on preserving scientific annotations.
37. So far hosting all infrastructure internally, using smaller datasets right now; scale as much as possible in-house.

Geodisy

38. Many researchers looking for spatial data, no easy way to do it now, and lack of accessibility in searching GIS.
39. Starting with DV to create a Google Map-like interface to search for datasets via geolocation.
40. Additional info on steps:
    a. Going to use existing tools connected via pipeline that queries DV repositories for geospatial (meta)data.

    b. Will harvest and clean data, and serve it to any instances (including GeoBlacklight) that can use it.
    c. Looking for any potential collaborations, overlaps
41. FAIR:
    a. Findability and Accessibility: crosswalking to existing standards, cleaning data.
    b. Interoperability: APIs to make data interoperable.
    c. Reusability: trying to make the data more accessible and findable.
42. Cleaning data - how difficult is that? Opportunity for machine learning?
    a. Very hard, DV is a self-deposit system so metadata quality varies. Some DVs that are mediated so quality is higher. Have a metadata developer to figure out how to clean it, so will be looking into approaches.
43. Can you explain what is meant by manuals for crosswalk?
    a. The project's metadata analyst will create manuals with appropriate metadata fields and controlled vocabularies for deposit.
44. Genomics - are there geo-based search interfaces for genome data, or would that be useful to that community?
    a. None that anyone could point to, but projects with a focus on population genetics could fund this feature useful.
45. Geoserver: provides necessary publishing to publish/share geospatial data over web, so people can query data.
46. Scope of project is DV only, but if project deliverables allow, may want to experiment with others.

MINTED

47. Working with real-time, and non-real time data, mixed data, heterogeneous types of data
48. Implementing data citations:
    a. Following 2016 RDA Data Citations WG guidelines.
49. Will have to add resolving landing pages for datasets, hook up DataCite and ORCID.
50. Metadata will be retrievable via a resolving web service and landing page.
51. A lot of interoperability happening with ISO and other standards that are being used.
52. Consideration: end users, how do they cite data?
53. Versioning
    a. Already keep track of agents that constitute change, need to make decisions about what triggers a new version.
    b. Keep track of when an agent is used? Or compare results? Two approaches that are being looked at.
54. Basic premise of RDA guidelines: query that can be local to the system and dataset DOI. Combination of both used to articulate what was used in research.
    a. Already store queries and sort records in reproducible way; time stamped query.
    b. Don't quite have uniqueness (e.g. two people asking for same exact query, they're stored twice).
    c. Have monthly automated data product comparisons again product standards.

55. Where in workflow does it make sense to mint a dataset?
56. Oceans 2.0 - data search interface and visualization tools already exist for system and web.
    a. Search info is stored by ONC.
57. Data search - ISO record generated on the fly during search, don't yet have a server to harvest from.
58. What third parties is ONC is hosting data for?
    a. Two kinds: third party maintains equipment on their own, ONC works to get data stream from them; ONC also maintains instruments while researchers operate them in field.
59. Granularity - do you ID relationships between data?
    a. Have fine scale granularity (specific fields) allows for associations/relationships between datasets, allows system to aggregate data.
60. DataCite being used a lot, some concern about gaps in support and sustainability?
    a. Discussion of DataCite as a core infrastructure.

Radiam

61. Goal is to come up with generic platform to organize research data - lots of platforms out there are domain-specific, less reusable for other disciplines.
62. Fewer DM solutions for data actively being collected, processed and analyzed than for completed data.
    a. Becoming less feasible to store large amounts of data in one place.
        i. EG Global Water Futures is the test case: Over a petabyte of 60 different types of data across 18 institutions.
            1. How do we accommodate this kind of data and get it organized?
            2. Need tools to manage it through entire RDM lifecycle - while it's active.
63. System diagram:
    a. Heart of radiam: research project (eg Global Water Futures).
    b. Central metadata index: querying.
    c. Data producers: where data currently lives, geographically distributed, different kinds of storage.
        i. One goal is have an agent able to harvest appropriate metadata for all of these for indexing.
        ii. API plugin that allows indexing in preexisting platform.
        iii. webGUI: data managers can interact, annotate, add domain-specific metadata to data, control permissions; researchers can query metadata, find what's out there, ask Radiam if it can broker transfer access to the data.
        iv. Can push a package of aggregated data that is suitable for ingest into a repository for publication
64. Want to build a platform that allows people to build more complicated tools on top.

65. Very fluid environment at active storage level - how are you handling versioning in this environment?
    a. Leaving data on other people's storage means they can make changes; versioning is a challenge; relying on platforms/tools to make note of changes in data.
    b. Dashboard to see change rates; hotspots of versioning could be interesting metrics.
66. Can you imagine an opportunity when someone wants to publish a dataset: the systek can automatically determine the data based on file type etc. and where it should go?
    a. Can imagine that, implementing that might be difficult

Research Portal for Secure Data Discovery, Access and Collaboration

67. A lot of programs have data modalities, but no centralized platform for them to upload and share data.
68. Opportunity to bring data together across different consortia and disciplines.
    a. Challenge: dealing with large scale research programs, setting up a central platform hard.
    b. Dealing with different types of data coming in - clinical, imaging, molecular, patient information.
69. Brain-CODE
    a. Working with partners to expand platform's services.
        i. Central platform deployed at Centre for Advanced Computing (Queen's U)
        ii. Different models for use and expansion of the platform
        iii. e.g. standalone instance deployed within a hospital (CAMH) vs. institutions or research programs using the central platform deployed at CAC
    b. Data capture tools for different kinds of data.
        i. e.g. REDCap, OpenClinica, XNAT, LabKey, etc.
        ii. Portal should integrate commonly used tools
    c. Continuously running pipelines to transform, clean, curate data.
    d. Other services - security and ID management, access, ethics, etc.
    e. Interfaces back through portal for researchers - dashboards for quality control, admin functions, study-specific ones; tools for analysis.
    f. Can link data in Brain-CODE to external sources through encrypted identifiers.
    g. Issue: needs context-dependent dashboard view of what tools matter to the user.
70. Various platforms also planned and deployed by Indoc in other jurisdictions and disease areas e.g. cancer, critical care, etc. - all requiring common data portal functionality
71. CANARIE focus:
    a. Customizability
        i. Hosting micro-sites within same installation.

        ii. Framework needed for building portals; currently evaluating existing CMS and framework options, want to avoid building everything from scratch
    b. Identity, authentication, authorization
        i. Layered on top of framework to provide unified identity and single sign-on across integrated tools, as well as user management features to manage accounts and access control
        ii. Focusing on internal ID services using FreeIPA, Keycloak and other open source services, branching out to other external authentication services down the road, e.g. ORCID
    c. Data access
        i. Data capture, analysis, vis tools; ideally integrated into portal.
        ii. Goal is to provide access to data via integrated data capture tools, as well as built-in data request and data query interfaces
    d. Collaboration
        i. Sharing, communicating within the platform, e.g. file sharing, messaging
        ii. Important for these features to be secure and fully integrated when working with sensitive/confidential information
    e. Focusing on first two right now, June for first release.
72. Platform presents so many different data sources - is there a data model?
    a. The data model is flat as possible. Treat each data source as its own dataset coming into platform; clinical data has more structured model; models not necessarily based on standards but looking into this to inform the models and tracking consent (e.g. BIDS, DATS, models for structuring ethics information).
73. Are you looking into dynamic consent?
    a. Initial requirement was fixed consent, but with the ability to withdraw and modify consent. Increasing ability to manage online consent a need and would like to incorporate that; data capture tools such as REDCap and OpenClinica provide eConsent functionality, so prefer to expose these rather than redevelop.

Islandora

74. Leveraged a lot of work done by SFU and the Islandora Community.
    a. A lot of islandora community work is modular and reusable.
    b. Slightly different approach with this new effort: previous 7x version supported researcher from planning to publishing and preserving.
    c. Knew researchers were using Dropbox and other storage and wanting to provide internal storage for active data.
75. FAIR
    a. Looking into minted DOIs for published datasets.
    b. Librarians can work with researchers to determine appropriate standards and implement them, as system can accommodate pretty much anything.
    c. Building in DataCite’s metadata schema as a discipline-neutral default

    d. Metadata and their identifiers are searchable and displayed; metadata will be registered and indexed with DataCite on DOI registration.
    e. Integration with ORCID and the use of ORCIDs within records
    f. REST endpoint
76. NDSF
    a. Look at Datacite specific module for platform.
    b. Endpoint for FRDR harvest.
    c. Islandora 8x leverages storage abstraction layer
        i. Abstraction layer is Flysystem and there are a number of adaptors available. An adaptor for Fedora has been created and is being tested.
            1. Eg. Openstack Swift is support so storage on Compute Canada or the ORLC could be possible.
        ii. Collab with DuraCloud instance?
    d. Looking at sharing ontologies, storage, etc. with repositories.
    e. Already produced a checksum checking service, bagit service modules for preservation.
    f. Tools are hooks to researchers to get them interested eg DMPs.
    g. Interested in API work that DMP folks are doing.
        i. Being able to push DMPs from UPEI into DMP Assistant is also an interest,
77. Goal: not everyone has a full team of devs to set up a repository; opportunity to talk to researchers at institution, position library as source of knowledge and expertise; capacity-building in institution.